Character-Sense Association and Compounding Template Similarity: Automatic Semantic Classification of Chinese Compounds

Chao-Jan Chen

LATTICE, University Paris VII, Paris, France

chen_chaojan@yahoo.com.tw

Abstract

This paper presents a character-based model of automatic sense determination for Chinese compounds. The model adopts a sense approximation approach using synonymous compounds retrieved by measuring similarity of semantic template in compounding. The similarity measure is derived from an association network among characters and senses, which is built from a formatted MRD. Adopting the taxonomy of CILIN, a system of deep semantic classification (at least to the small classes) for V-V compounds is implemented and evaluated to test the model. The experiment reports a high precision rate (about 38% in outside test and 61% in inside test) against the baseline one (about 18%).

1. Introduction

Sense tagging is an important task in NLP. It is supposed to provide semantic information useful to the application tasks like IR and MT. As generally acknowledged, sense tagging is to assign a certain sense to a word in a certain context by using a semantic lexicon (Yarowsky, 1992, Wilks and Stevenson, 1997). In addition to word sense disambiguation (WSD) for known words, sense determination for words unknown to the lexicon poses another challenge in sense tagging. This is especially the case in NLP of Chinese, a language rich in compound words. According to the data in (Chen and Lin, 2000), about 5.51% of unknown words is encountered in their sense-tagging task of Chinese corpus. Instead of proper names, the cross-linguistically most common type of unknown words, compound words constitute the majority of unknown words in Chinese text. According to Chen and Chen (2000), the three most dominant types of Chinese unknown words are: compound nouns (about 51%), compound verbs (about 34%), and proper names (about 15%). While the identification and classification of proper names is an issue already well discussed in Chinese NLP researches, the sense determination of unknown compounds remains a subject relatively less tackled.

1.1 Shallow vs. Deep Classification

While *word sense* might be conceptually vague and controversial in linguistics and difficult to define (Manning and Schütze, 1999), *sense tag* is more concrete and can be defined according to the specific need of the NLP tasks in question. For example, in a task of semantic tagging or classification, *sense tag* can be the semantic class from a thesaurus. Or otherwise, in a task of machine translation, the equivalent foreign word from a bilingual dictionary can be chosen as *sense tag*. In this paper, it is the *sense tag* so defined that is meant by the term *sense*. The notion *sense determination* then refers to the assignment of sense tag to a word without using contextual information. It is so called to be distinguished from *sense tagging*, which requires contextual information. Under such a definition, *semantic classification* can be regarded as a case of sense determination using the taxonomy of a certain thesaurus, in which a *semantic class* is a *sense tag*.

According to Wilks and Stevenson (1997), a task assigning *broad* sense tags like HUMAN, ANIMATE in WordNet is referred to as *semantic tagging*, different from *sense tagging*, which assigns more *particular* sense tags. In fact, a similar distinction can also be made for semantic classification according to the target level of the semantic classes in the taxonomy tree: a task aiming at the top-level classes can be called *shallow* semantic classification (like Lua, 1997), while a task aiming at the bottom-level classes can be called a *deep* semantic classification[1] (like Chen and Chen, 2000). Since many top-level semantic classes, like TIME, SPACE, QUALITY, ACTION, etc., are often already reflected in the syntactic information, a shallow semantic classification does not actually provide much semantic information independent of syntactic tagging. It is therefore the deep semantic classification that the paper is concerned about.

[1] Take the word 攻擊(‘attack’) for example. According to CILIN (a thesaurus widely used in Chinese semantic classification, see 3.1), it can be classified to shallow-levels as major class H (ACTIVITY) or as medium class Hb (MILITARY ACTIVITY). It can also be classified to deep-levels as small class Hb03 (specific military operations: ATTACK, RESIST, and COUNTERATTACK) or as subclass Hb031 (ATTACK).

1.2 Previous Researches

In the previous researches of automatic semantic classification of Chinese compounds, compounds are generally presupposed to be endocentric, composed of a head and a modifier. Determining the class of the head is therefore determining the class of the target compound (Lua, 1997, Chen and Chen, 2000). This head-determination approach has two advantages: (1) it is simple and easy to implement (2) it works effectively for compound nouns, the dominant type of compounds, since most of them are head-final endocentric words.[2] However, there exist considerable exocentric compounds, for which such a simple algorithm does not work successfully. It is especially the case for compound verbs like V-Vs[3]. For example, 打死 is a V-V compound meaning 'to kill by beating'. Obviously, neither the sense of 打 ('beat') nor that of 死('die') is appropriate to be assigned to the compound 打死 as the sense of 車 ('car') can be assigned to 電車('tram', literally 'electricity-car') as a general meaning.

A second problem encountered in compound semantic classification is that there are considerable *out-of-coverage morphemes*, which are not listed in the lexicon, as remarked in (Chen and Chen, 2000). Moreover, even a morpheme is listed, the given senses are not necessarily appropriate to the task. For example, in the search of compound morphological rules in (Chen and Chen, 1998), some appropriate senses of morphemes have to be added manually to facilitate the task. Obviously this causes a great difficulty to an automatic task, especially to the example-based models which rely on the similarity measurement of the modifier morphemes to disambiguate the head senses (Chen and Chen, 1998, 2000). An alternative approach is thus needed to solve the problems of exocentric compounds and lexicon incompleteness.

Therefore in this paper I will present a non head-oriented model of Chinese compound sense determination, in which lexicon incompleteness will be overcome by exploring the association between characters and senses in a MRD. The sense of an unknown compound can be approximated by retrieved synonyms. Its sense tag can be assigned according to a certain MRD. This model facilitates an automatic system of deep semantic classification for unknown compounds. In this paper, a system for V-V compounds is implemented and evaluated. The model can however be extended to handle general Chinese compounds, like V-N and N-N, as well.

2. Compound Sense Determination

2.1 Compounding Semantic Templates

Most of the Chinese compounds are composed of two constituents, which can be bound morphemes of one character or free words of one or more characters. The two-character compound is a most representative type because its components can be bound morphemes as well as free words. The handling of two-character compounds becomes therefore the focus in this paper.

As in general Chinese compounding, a two-character compound is usually semantically compositional, with each character conveying a certain sense. The principle of semantic composition implies that under each compound lies a semantic pattern, which can be represented as the combination of the sense tags of the two component characters. The combination pattern is referred to as *compounding semantic template* (denoted by S-template) in this paper; compounds of the same S-template are then referred to as *template-similar* (denoted by T-similar). Since T-similar compounds are alike in their semantic compositions, they are supposed to possess roughly the same meaning and to be put under a considerably fine-grained semantic class. Take the compound verb 打破 for example. This compound suggests the existence of a *S-template* of HIT-BROKEN, as the senses of the two component characters 打 and 破 are respectively 'hit' and 'broken'. The S-template HIT-BROKEN refers to a complex event schema [to make something BROKEN by HITting]. This S-template can also be found in many other compounds with a similar meaning:打碎,擊碎,敲破,敲碎…etc. Obviously such T-similar words can make a good set of examples for the example-based approach to the sense determination, if an effective measure of word similarity is available for their retrieval.

2.2 Compound Similarity

As a critical technique, word similarity is generally used in the example-based models of semantic classification. The measure of word similarity can be

[2] Though a compound noun and its head are strictly speaking in a hyponym relation, they are usually categorized as members of the same class. For example, in CILIN,車('car', 'vehicle') and most of the compounds X-車 are put under the same class Bo21 (VEHICLES), where X can be a morpheme designating the energy source (like horse, cow, electricity) or the load content (like passenger, merchandise).

[3] An introspection on the two-character verbs in CILIN shows that about 48% of them are semantically exocentric, which means the semantic class of a compound X-Y in CILIN is equal neither to that of X nor to that of Y. As to the endocentric V_1-V_2, V_1 and V2 are about equally likely to be the head of a compound verb according to the introspection.

divided into two major approaches: taxonomy-based lexical approach (Resnik 1995, Lin 1998a, Chen and Chen 1998) and context-based syntactic approach (Lin 1998b,Chen and You 2002), which is not the concern in this context-free model. However, two problems arise here for the taxonomy-based lexical approach. First, such similarity measures risk the failure to capture the similarity among some semantically highly related words, if they happen to be put under classes distant from each other according to a specific ontology[4]. Second, as mentioned, the appropriate senses of some characters just cannot be found in the thesaurus. One major reason why dictionaries do not include certain character senses is that many of such characters are used in contemporary Chinese only as bound morphemes not as free words, when the senses in question are involved. However, such senses could be kept in the compounds in the lexicon, so they might be covert but not inextricable.

To remedy the effects of such lexicon incompleteness, I propose an approach to retrieve the *latent* senses[5] of characters and the *latent* synonymy among characters by exploring association among characters and senses. The idea is that if a character C appears in a compound W, then according to semantic composition, the sense of C must somehow contributes to S, the sense of W. Therefore the association strength between character C and sense S in a MRD is supposed to reflect the potentiality of S to be a sense of C. By transitivity, such association between characters and senses allows to capture association among characters. A new way to measure word similarity of two compounds can be thus derived based on the association strength of the corresponding component characters. This measure actually reflects the S-template similarity between two compounds and can be used to retrieve for a compound its T-similar words, which are potentially synonymous.

2.3 Synonyms and Sense Approximation

The acquisition of synonyms plays an important role in the sense determination of a word. When a native-speaker is capable of giving synonyms to a word, he is considered to understand the meaning of that word. In fact, such a way of sense capturing is also reflected in how the senses of words can be explained in many dictionaries[6]. Moreover, as some researches propose, synonyms can be used to construct the semantic space for a given word (Ploux and Victorri, 1998, Ploux and Ji, 2003). In such a semantic space, each synonym with different nuance occupies a certain area. As visually reflected in this approach, retrieving a proper set of its synonyms means the ability to well capture the senses of a word. In fact, my model of automatic sense determination for a compound is exactly built upon the retrieval of its near synonyms, the T-similar compounds as previously described.

2.4 Model Representation

With a S-template similarity measure, one can retrieve, for a given compound, its potential synonymous T-similar compounds. Then the sense tags of the retrieved compounds can be used to determine the sense tag of the target compound. The model of compound sense determination can be thus composed of two modules, as illustrated in Fig.1.

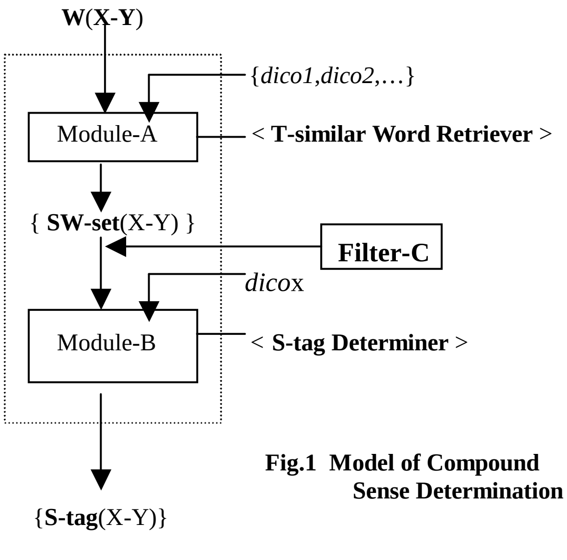

Fig.1 Model of Compound Sense Determination

Module-A (<T-similar Word Retriever>) is to find the potential synonyms ({**SW-set**(X-Y)}) of a given compound (X-Y) by using association information provided from dicos {dico1, dico2,…}. Module-B (<S-tag Determiner>) is to obtain the most likely

[4] Take an example in CILIN (a Chinese thesaurus, see 3.1). KILL(殺死), BUTCHER(屠宰), and EXCUTE(處決) are three concepts all meaning ‘cause to die’. However, the words expressing these three ideas are respectively put under small classes Hn05, Hd28, and Hm10, respectively under medium class Hn: *Criminal Activities*(惡行), class Hd: *Economical Production Activities*(生產), and class Hm: *Security and Justice Activities*(公安, 司法). We wonder if any measurement based on that hierarchy can capture the similarity among the words situated in these three small classes in CILIN, for those words share only a common major class H, denoting vaguely *Activities*, which includes 296 small classes and 836 subclasses.

[5] Here the term *latent* is used only to mean ‘hidden, potential, and waiting to be discovered’. It has nothing to do with the LSI techniques, though they both evoke the same meaning of *latent*.

[6] Especially in Chinese dictionaries, it is often the case that several synonymous words are given as explanation to the meaning of a word, especially when it is a compound verb.

sense tags ({**S-tag**(X-Y)}) according to dicox for the target word by using the output of Module-A. The component filter-C is optional, which passes only the T-similar words with the same syntactic category as the target compound, if it is already known. In fact, a system of semantic classification can be so created by choosing dico2 as dicox and the S-tag is then the semantic class in CILIN (as in section 4).

3 Character-Sense Association Network

Before exploring the critical measurement of association among characters and senses needed in the model, I have to briefly present the lexical sources in use and to define the idealized dictionary format adopted in this task.

3.1 Lexical Sources

The lexical sources used to implement my system include:

(1) **Sinica Corpus**: a balanced Chinese corpus with 5 million words segmented and tagged with syntactic categories. (Huang et al., 1995)

(2) **HowNet**: an on-line Chinese-English bilingual lexical resource created by Dong. It is used in this paper as a Chinese-English dictionary registering about 51,600 Chinese words, each assigned with its equivalent English words and its POS. (http://www.keenage.com/)

(3) **CILIN**: a Chinese thesaurus collecting about 53,200 words. CILIN classifies its lexicon in a four-level hierarchy according to different semantic granularities: 12 major classes (level-1), 95 medium classes (level-2), 1428 small classes (level-3), and 3924 subclasses (level-4). The words in the same small class can be regarded as semantically similar, but only the words in the same subclasses can be surely regarded as synonyms[7].(Mei et al., 1984)

3.2 Idealized Dictionary Format (*dico*)

The idealized dictionary, denoted as ***dico***, is actually a formatted MRD defined as follows:

A *dico* is a set of <*W*-*S*> correspondence pairs,
where *W* is a word, and *S* is a sense tag. **(1)**

In the system implementation in this paper, two *dico*s are converted respectively from HowNet and CILIN for the calculation of the association measures among characters and sense tags with different types of sense tags adopted. For HowNet, the English equivalent words are used as sense tags to form dico1. For CILIN, the subclasses are used as sense tags to form dico2.

3.3 Character-Sense Association

All the semantic information provided by a *dico*, as defined in (1), can be in fact represented as a network with links between two domains: *W* domain (words) and *S* domain (sense tags). In such a viewpoint, polysemy is then a one-to-many mapping from W to S, while synonymy a one-to-many mapping from S to W. If we further link a component character C of a word W to one of the S linked to W, such a C-S link might intuitively reflect a potential sense S for the character C, probably a *latent* sense of C, as previously described in section 2.2. We can use a statistical association measure, like MI or χ^2, to extract such C-S links. The statistically extracted C-S association can then lead to the finding of latent senses for a character. The revelation of a latent character-sense association will further lead to the retrieval of new synonymy relation between characters. Symmetrically, the revelation of a latent character-sense association will also lead to the retrieval of the potential polysemy of a character. As illustrated in the Z-diagram below, supposed that *C*1 is already associated to *S*1 and *C*2 to *S*2, the retrieval of *latent sense S*1 to *C*2 will, meanwhile, lead to the finding of an association between *C*1 and *C*2 (*latent synonymy*), and an association between *S*1 and *S*2 (*latent polysemy*).

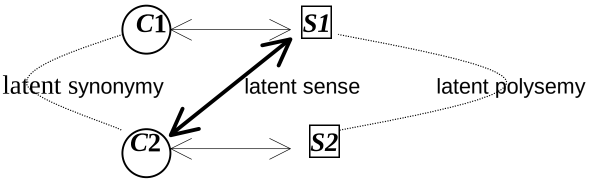

Fig. 2 Z-diagram of C-S links

The *directed* association measure from a character to a sense, denoted as CS-asso(C_i,S_j), can be defined as follows:

$$\alpha(C_i, S_j) = [\ \mathrm{freq}(C_i,S_j)^2 / (\ \mathrm{freq}(C_i)+\mathrm{freq}(S_j)\ )\ ] \wedge 0.5$$

$$\text{CS-asso}\ (C_i, S_j) = \alpha\ (C_i,S_j)\ /\ \mathrm{Max}_k \{\ \alpha\ (C_i,S_k)\ \} \qquad (2)$$

where freq(C_i,S_j) is the number of the words in the MRD that contain character C_i and is tagged with sense S_j, while freq(C_i) is the number of words

[7] Take two verbs 買('to buy') and 賣('to sell') as examples to demonstrate the taxonomy of CILIN. Both of the two verbs are grouped in the small class He03 (commercial trade), which is under the major class H (activities) and the medium class He (economic activities). However, the two antonyms are put under two different subclasses, respectively He031 (buying) and He032 **(**selling).

containing character C_i, and freq(S_j) the number of words tagged with sense S_j.[8]Likewise, the *directed* association measure from a sense to a character, denoted as SC-asso(S_i,C_j), can be defined as follows[9]:

$$\alpha\,(S_i,C_j) = [\ \mathrm{freq}(S_i,C_j)^2/(\ \mathrm{freq}(S_i)+\mathrm{freq}(C_j)\ )\ ]\ \text{^}0.5$$
$$\text{SC-asso}\,(S_i,C_j) = \alpha\,(S_i,C_j)\ /\ \mathrm{Max}_k\,\{\ \alpha\,(S_i\ ,\ C_k)\ \}, \qquad (3)$$

Consequently, by link of a C_i-S_j-C_k chain (a latent synonymy), the directed association measure for a character C_i to another character C_k is defined as a combination of two types of directed association measures, the maximal association measure CC-asso1(C_i ,C_k) and the over-all association measure CC-asso2(C_i ,C_k), with respective weights of 1-ω and ω (the value ω is by default set at 0.5).

$$\text{asso-chain}(C_i,S_j,C_k) = (\text{asso}\,(C_i,S_j)\ *\ \text{asso}\,(S_j,C_k)\ )\ \text{^}\ 0.5$$
$$f1\,(C_i,C_k) = \mathrm{Max}_j\,\{\text{asso-chain}\,(C_i,S_j,C_k)\ \}$$
$$\text{CC-asso1}(C_i,C_k) = f1\,(C_i,C_k)\ /\ \mathrm{Max}_m\,\{\ f1\,(C_i,C_m)\ \}$$
$$f2\,(C_i,C_k) = \Sigma_j\ \text{asso-chain}(C_i,S_j,C_k)$$
$$\text{CC-asso2}(C_i,C_k) = f2\,(C_i,C_k)\ /\ \mathrm{Max}_m\,\{\ f2\,(C_i,C_m)\ \}$$
$$\text{CC-asso} = (1-\omega)\ *\ \text{CC-asso1} + \omega\ *\ \text{CC-asso2} \qquad (4)$$

3.4 S-Template Similarity Measure

Supposed that W_i(C_{i1}-C_{i2}) and W_j(C_{j1}-C_{j2}) are both two-character compounds, a measure of word-word directed association (denoted as WW-asso) from W_i to W_j can be defined based on the CC-asso between their corresponding component characters:

$$\beta\,(W_i,W_j) = \{\ \text{CC-asso}(C_{i1},C_{j1})\ *\ \text{CC-asso}(C_{i2},C_{j2})\ \}\ \text{^}\ 0.5$$
$$\text{WW-asso}(W_i,W_j) = \beta\,(W_i,W_j)\ /\ \mathrm{Max}_k\,\{\ \beta\,(W_i,W_k)\ \} \qquad (5)$$

Since the corresponding characters of two T-similar compounds must share the same sense tags and thus have strong CC-asso, the measure WW-asso(W_i,W_j) indicates, in fact, how T-similar for a compound W_j to a target W_i, compared with other compounds. WW-asso(W_i,W_j) is therefore taken as the measure of S-template similarity (denoted as T-similarity).

Applying the S-template similarity measure in (5), now the T-similar Word Retriever (<TWR>) can give for a compound X-Y the list of its most T-similar compounds from the corpus and their T-similarity scores. As to the <S-tag Determiner>, it receives as input the output T-similar words from <TWR>. Among the input T-similar words, the ones known to $dico_x$, are picked out and their sense tags (S-tag) with the T-similarity scores (WW-asso) are used, as in the formula (6), to calculate the likelihood score Λ for a compound ***V*-*V*$_i$** to possess a certain **S-tag$_j$**. Therefore a set of ranked possible semantic classes for the compound X-Y can be given ({**S-tag**(X-Y)}).

$$\lambda(V\text{-}V_i,\ \text{S-tag}_j) = \Sigma_j\ \text{WW-asso}\ (V\text{-}V_i,\ SW_k) \qquad (6)$$

,where SW_k is a known word in $dico_x$
and $S\text{-}tag_j$ is one of the S-tages to SW_k

$$\Lambda(V\text{-}Vi,S\text{-}tag_j)=\lambda(V\text{-}V_i,S\text{-}tag_j)/\mathrm{Max}_n\,\{\ \lambda(V\text{-}V_i,\ S\text{-}tag_n)\ \}$$

4. System Implementation

4.1 Classification for V-V Compounds

Based on the model proposed, a system of semantic classification can be implemented for two-character V-V compound verbs by using dico2 as the dicox in the Module-B (the S-tag now is the semantic class in CILIN). The V-V compounds are chosen as subjects in this system because the choice can best distinguish the present model from the previous head-orientated approaches. As the involvement of only V characters make training data homogeneous, it simplifies the association network and reduces largely the computational complexity. However, the partial system for V-V compounds can be easily extended to handle V-N compounds and N-N compounds as well when the character-sense association network for N characters is established.

Since only the V characters are involved, a subset of <W-S> pairs of dico1 (HowNet) and dico2 (CILIN) is extracted to calculate the association measures and then the T-similarity measure. The subset contains only the <W-S> pairs whose W are one-character or two-character verbs. In CILIN the verbs are put under the major classes from E to J, designating the concepts of attributes (E), actions (F), mental activities (G), activities (H), physical states (I), and relations (J). By choosing only the words in the above 6 major classes, the nominal senses of characters (A: human, B: concrete object, C: time and space, D: abstract object) are supposed to be excluded. Besides, the occurrence frequency of a character in a mono-character word will be double weighted, since in this case the word sense is surely contributed by that character alone.

Let us take the V-V compound 捕獲 (‘to catch by hunting’, literally ‘hunt-catch’) for example to see how the model operates. Based on the

[8] The formula α in (2) is actually a simplified approximation to the χ^2 -test measure by supposing that freq(C,S) is much smaller than freq(C) and freq(S). In fact, MI (mutual information) is another association measure frequently used in Chinese NLP. For example, it is successfully used for the character-POS association measure in the task of syntactical classification for Chinese unknown words (Chen et al., 1997). However, a heuristic evaluation on some randomly picked examples shows that it seems to be outperformed by the χ^2 measure in this task.

[9] It must be noted that the measures of directed association (2) and (3) are asymmetric in that they give different values for the association from *C*i to *S*j and for the one from *S*j to *C*i because their normalization factors are not the same. That is why the notion *directed* is added here to point out the asymmetry.

association network created from HowNet, the characters associated to 捕 and 獲 are listed in List 1 and List 2 (only the 10 top ranked are listed here), the 20 top ranked T-similar compounds of 捕獲 are listed in List 3 with their similarity scores, syntactic categories and semantic classes, if they are known in CILIN. Among the 20 T-similar compounds retrieved, 10 of them (the grayed ones) can be found in CILIN; 9 of them (the framed ones) can be considered as good synonyms of 捕獲, while other 7 (the starred ones) considered semantically really close. In this particular example, 80% (16/20) of the T-similar compounds can be considered as at least near synonymous, while 50%(8/16) of them can be actually found in CILIN to serve the automatic semantic classification.

捕 1.0000	抓 0.8832
緝 0.9634	擒 0.8673
捉 0.9165	佔 0.8629
逮 0.9073	獲 0.8558
拿 0.9022	奪 0.8306

List 1

獲 1.0000	賺 0.8694
得 0.9402	俘 0.8687
擒 0.9146	捉 0.8641
拿 0.9076	捕 0.8632
收 0.9034	繳 0.8614

List 2

捕獲 1.0000 VC H m051	緝捕* 0.8316 VC Hm051
緝獲 0.9634 VC H m051	佔得 0.8113 VC
捕得 0.9402 VJ	獲得* 0.8046 VJ Je121
逮獲 0.9073 VC	獵獲 0.8036 VC
拿獲 0.902 2 VC Hm051	捕抓* 0.7970 VC
緝拿* 0.8744 VC Hm051	擒拿 0.7872 VC Hb141
擒獲 0.8673 VC	攫獲 0.7853 VC
捕捉* 0.8641 VC	逮捕* 0.7832 VC Hm051
捕到 0.8380 VC	奪得 0.7809 VC Je121
捉拿* 0.8318 VC Hm051	扣得 0.7790 VC

List 3

Applying the formula for the likelihood score of semantic class determination in (6), we have the 4 top ranked semantic classes for 捕獲 predicted by the system as follows:

(1) Hm051 (逮捕 'arrest')
(2) Je121 (取得 'acquire')
(3) Hb121 (攻佔 'attack and occupy')
(4) Hb141 (俘虜 'capture as war prisoner')

In this case, the standard answer of class Hm051 for the compound 捕獲 is ranked as the first candidate, while the second ranked candidate class Je121 ('acquire') is also reasonable, which can be considered rather correct in a certain way by human judgment. In fact, according to the native speaker's instinct, the 4th ranked candidate class Hb141 ('capture') is also quite suitable to the meaning of the verb 捕獲, though that is not what it is classified in CILIN. However, to avoid the subjective interference of human judgment and particularly to make the evaluation task automatic, the evaluation in the following sections will be made by machine only according to the standard classification in CILIN.

4.2 Experiment Results

For evaluating the performance of the system, 500 V-V compounds are randomly picked out from CILIN to form the test set. Two modes of evaluation experiments are carried out: both modes adopt dico2 (CILIN) in Module-B (dicox=dioc2) to determine semantic classes, while the inside-test mode uses dico2 (CILIN) in Module-A and the outside-test mode uses dico1 (HowNet) in Module-A, to obtain association network and retrieve the T-similar words. To make the test compounds *unknown* to the model, the semantic classes of the test compounds have to be *invisible* to CILIN, while the *invisibility* should not undermine the training of the association network in Module-A. The effect is done by dynamically withdrawing a word from dico2 in Module-B each time when it is in test. Two ways of evaluation can be made: by verifying the answer to the level of small class (level-3) and to the level of subclasses (level-4). The accuracy is calculated by verifying if the correct answer or one of the correct answers (if V-V is polysemous) according to CILIN can be found in the first n ranked semantic classes predicted by the system. The performance of a random head-picking model is offered as the baseline. In this baseline model, one of the semantic classes of X and Y is randomly chosen as the semantic class of the compound X-Y.

n	Level-3(Small Class)			Level-4(Subclass)		
	outside	inside	Baseline	outside	inside	Baseline
1	**39.80%**	**61.60%**	**18.83%**	**36.60%**	**60.40%**	**17.34%**
2	56.80%	76.00%	31.40%	52.80%	74.40%	29.12%
3	64.40%	83.80%	40.21%	59.80%	80.80%	37.54%

Table 1. Performance for 500 V-V compounds

The results in Table 1 show that the system achieves a precision rate of 60.40% for inside test and 36.60% for outside test in level-4 classification against the baseline one of 17.34%. Not to our surprise, the performance of classification to level-3, a slightly shallower level, is slightly better: 61.60% for inside test and 39.80% for outside test. Table 1 also shows that the system can achieve a correction rate of 59.8% (outside) and 80.80% (inside) for including the correct answer in the first 3 ranked candidate classes in level-4, 64.40% (outside) and 83.80% (inside) in level-3, all much better than the baseline ones, 37.54% and 40.21%.

4.3 A Pseudo-WSD Problem

If the correct semantic class can be found in a limited number of candidates, context information can be used to help determine which candidate is more likely to be the proper one, just as a WSD task does. Take again the example of the compound 捕獲 in section 4.1, which the system classifies most likely as: ‘**arrest**’, ‘**acquire**’ and ‘**attack-occupy**’. Obviously the verbs in the three classes should take different stereotypes of objects: respectively **person**, **thing**, and **place**. Therefore it is not difficult to determine the correct semantic class of the verb in question by using context information, in this case the type of the object. Through this example, we can see that the high inclusion rate of the correct answer in the top ranked classes has in fact a great significance: the ranking of the top candidates can be further adjusted and eventually ameliorated by context information, and thus the task of class determination can become a pseudo-WSD problem, in which domain various techniques are well available (Manning and Schutze, 1999). The performance of the present non-contextual system of automatic semantic classification is then expected to be improvable with the eventual help of a good context-sensitive WSD system, though it is out of the scope in this paper. Therefore the correct inclusion rate of top n ranked classes is also the concern of this paper.

4.4 Endocentric vs. Exocentric Compounds

Table 2 shows the performance of the system on the endocentric compounds (with heads) and on the exocentric ones (without heads) in level-3. Among the 500 V-V compounds, the endocentric V-V compounds have much higher precision rates than the exocentric ones. But even for the exocentric compounds, the precision rate of the system is 49.28% in inside test and 27.05% in outside test, while the correct inclusion rate of top 3 ranked classes achieves 74.64% in inside test and 51.69% in outside test. Such a performance is in fact rather encouraging since it shows that this model has overcome the inherent difficulty met by a head-oriented approach.

n	Outside		inside	
	+Head	-Head	+Head	-Head
1	**48.81%**	**27.05%**	**70.69%**	**49.28%**
2	68.26%	40.58%	84.83%	64.11%
3	73.38%	51.69%	90.69%	74.64%

Table 2. Level-3 performance for [+/- Head] V-V

4.5 Syntactic Category Filter

To test the function of the Filter-C in the model, two sets of 500 V-V compounds are randomly picked out from verbs of category VC corpus, and from verbs of category VA in Sinica Corpus.[10]. Table 3 and 4 show the performance of the system on the two kinds of verbs when evaluated to level-3. The results show that the system using the syntactic category filter (+SCF) performs slightly better than that without using the filter (-SCF) only in the precision of first ranked class in the outside test. Beside that, the use of the syntactic category filter generally undermines the performance of the system. Such a result might be explained by the fact that synonymous words in CILIN are not necessarily of the same syntactic category; it also suggests that for the entire model recall is perhaps more important than precision in Module-A.

n	outside		Inside		Baseline
	+SCF	-SCF	+SCF	-SCF	
1	**49.60%**	**47.60%**	**64.40%**	**67.20%**	**22.90%**
2	63.20%	64.00%	76.40%	78.40%	39.74%
3	70.00%	73.60%	84.40%	84.80%	50.27%

Table 3. Level-3 performance for V-V of category VC

n	outside		inside		Baseline
	+SCF	-SCF	+SCF	-SCF	
1	**41.00%**	**38.60%**	**52.00%**	**58.60%**	**15.90%**
2	52.20%	49.60%	67.40%	73.80%	26.84%
3	55.80%	55.00%	73.00%	80.20%	34.61%

Table 4. Level-3 performance for V-V of category VA

4.6 Classification Errors

An examination of the bad performing cases suggests that there are three major sources of erroneous classification in the experiments. (1) Some test compounds are just idiomatic or non semantic compositional. Naturally, it is highly difficult, if not impossible, to correctly predict their semantic classes. (2) Some compounds are from unproductive S-templates, which causes the example sparseness of the T-similar compounds. The scarcity of examples will easily lead to a poor determination result caused by a low noise tolerance of occasional bad examples. (3) Some classifications predicted by the system are reasonable to native speakers, but happen not to be the case in CILIN as the standard answers.

5. Conclusions and Further Remarks

In this paper I have proposed a character-based model of sense determination for Chinese

[10] VC (transitive action/activity) and VA (intransitive action/activity) are the two most dominant types of two-character verbs in the corpus, occupying respectively 44% and 27% around. Here the statistics does not include the VH (intransitive state) verbs, because they generally correspond to the adjectives in English, and in deed they are categorized as adjective in HowNet.

compounds using compounding template similarity. Based on this model, a system of deep semantic classification for V-V compounds is implemented, which classifies compounds according to the taxonomy of CILIN to its deep-level (level-3 and level-4) classes. The evaluation experiment reports a fairly satisfactory precision rate of the first ranked predicted semantic class (about 38% in outside test and 61% in inside test) against the baseline one (about 18%). The results also show a high inclusion rate of correct answer in the top3 ranked classes, which suggests that in the future the present non-contextual system can cooperate with a WSD module using context information. Though the model is only tested on a partial system for V-V compounds, it can be extended to work for general compounds, like V-N and N-N, with the association network further established for N characters.

The model proposed in this paper has the following advantages: (1) It proposes a similarity measure of compounding template to retrieve potential synonyms for sense approximation, which avoids the inherent difficulty of head determination in a head-oriented approach and is thus capable of handling exocentric compounds. (2) It establishes a network of character-sense association, which allows the discovery of latent senses of characters, latent synonymy, and latent polysemy, thus remedying the incompleteness effect of the MRD in use. (3) It can carry out deep semantic classification, not just shallow classification assigning general and vague categories. (4) It requires only a simple format of idealized dictionary, which facilitates the conversion from a general MRD and allows an easy enhancement of the system by adding a new MRD.

However, as can be remarked in the discussion of classification errors, the performance of the model relies much on the productivity of compounding semantic templates of the target compounds. To correctly predict the semantic class of a compound with an unproductive semantic template is no doubt very difficult due to a sparse existence of the T-similar compounds. How to remedy such an effect is thus a challenging task in the future. In addition, how to generalize the present character-based model to make it applicable to compounds with multi-character component morphemes will be another essential task to undertake. Besides, a task of automatic lexical translation for Chinese unknown compounds will also be carried out in the future. The task can be executed under the very same structure of the present model, since the only difference will be the change of working *dico*x (from dico2 to dico1) in the Module-B. A pilot experiment has already shown encouraging results.

References

Chao-Jan Chen, Ming-hong Bai, and Keh-Jiann Chen. 1997. Category Guessing for Chinese Unknown Words. In *Proceedings of the Natural Language Processing Pacific Rim Symposium* 1997, pages 35-40.

Hsin-Hsi Chen and Chi-Ching Lin. 2000. Sense-tagging Chinese corpus. In *Proceedings of ACL-2000 workshop on Chinese Language Processing*, pages 7-14.

Keh-Jiann Chen and Chao-Jan Chen. 1998. A Corpus Based Study on Computational Morphology for Mandarin Chinese. In Quantitative and Computational Studies on the Chinese Language, pages 283-306.

Keh-Jiann Chen and Chao-Jan Chen. 2000. Automatic Semantic Classification for Chinese Unknown Compound Nouns, In *Proceedings of Coling*-2000, pages 173-179.

Keh-Jiann Chen and Jia-Ming You. 2002. A Study on Word Similarity Using Context Vector Models, *Computational Linguistics & Chinese Language Processing*, 8(2):37-58.

Chu-Ren Huang et al. 1995. The Introduction of Sinica Corpus. In *Proceedings of ROCLING* VIII, pages 81-99.

Dekang Lin. 1998a. An Information-Theoretic Definition of Similarity, In *Proceedings of International Conference on Machine Learning*, pages 296-304

Dekang Lin. 1998b. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL-98*, pages 768–774.

Kim-Teng Lua. 1997. Prediction of Meaning of Bi-syllabic Chinese Compound Words Using Back Propagation Neural Network. In *Computer Processing of Oriental Languages*, 11(2):133-144.

Christopher Manning and Hinrich Schütze. 1999. Fondations of Statistical Natural Language Processing, MIT Press.

Jia-Ju Mei et al. 1984. TonYiCi CiLin – thesaurus of Chinese words (同義詞詞林), Shangwu Yinshuguan (商務印書館香港分館), Hong Kong.

Sabine Ploux and Bernard Victorri. 1998. Construction d'espace sémantique à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, 39(1):161-182.

Sabine Ploux and Hyungsuk Ji. 2003. A Model for Matching Semantic Maps Between Languages (French/English, English/French). *Computational Linguistics*, 29(2):155-178.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (IJCAI), pages 448-453.

Yorick Wilks and Mark Stevenson. 1997. Sense Tagging: Semantic Tagging with a Lexicon. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 47-51.

David Yarowsky. 1992. Word-Sense Disambiguation Using Statistical Models of Rogets Categories Trained on Large Corpora", In *Proceedings of COLING-92*, pages 454-460.